

Citation for published version:

Patel, M 2010, 'Integrated Research Data Management in the Structural Sciences: Scaling up to integrated research data management', Paper presented at Scaling Up to Integrated Research Data Management, I2S2 Workshop, IDCC 2010, Chicago, 6/12/10.

Publication date:
2010

Document Version
Early version, also known as pre-print

[Link to publication](#)

Publisher Rights
CC BY-SA

University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



Infrastructure for Integration in Structural Sciences

Manjula Patel

Scaling-up to Integrated Research Data Management Workshop
6th International Digital Curation Conference

Holiday Inn, Mart Plaza

Chicago, Illinois

6-8th December, 2010



This work is licensed under a Creative Commons Licence: Attribution-ShareAlike 3.0
<http://creativecommons.org/licenses/by-sa/3.0/>



Infrastructure for
Integration in Structural Sciences

Outline

- I2S2 Project overview and objectives
- Research Data & Infrastructure
- Requirements analysis
- A Scientific Research Activity Lifecycle Model
- An integrated information model
- Use cases
- Cost-Benefits Analysis



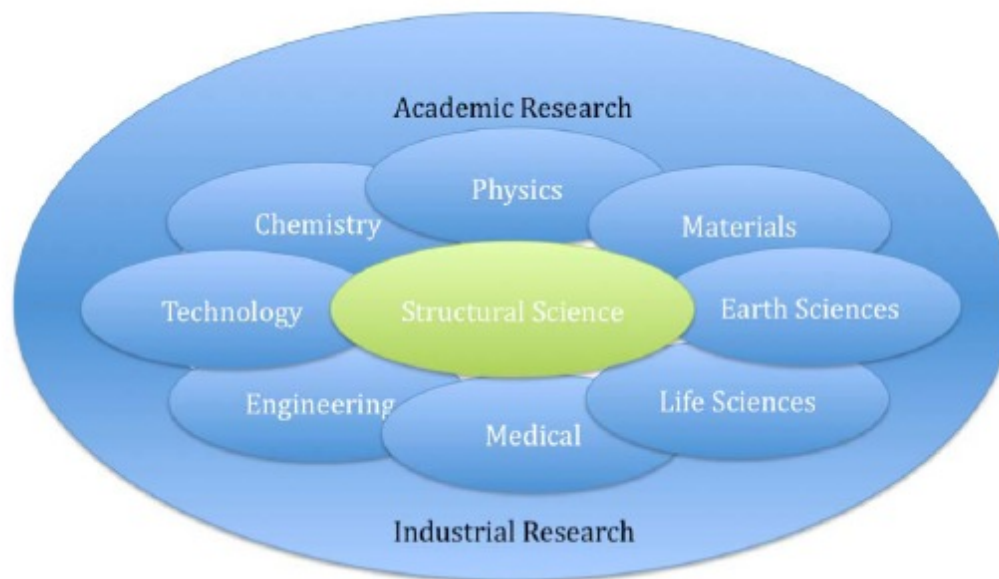
Diamond Light Source (DLS),
Science & Technology Facilities Council, UK



Infrastructure for
Integration in Structural Sciences

I2S2 Project Overview

- Understand and **identify requirements for a data-driven research infrastructure** in the Structural Sciences
 - Examine localised data management practices
 - Investigate data management infrastructure in large centralised facilities
- Show how effective cross-institutional research data management can **increase efficiency and improve the quality of research**





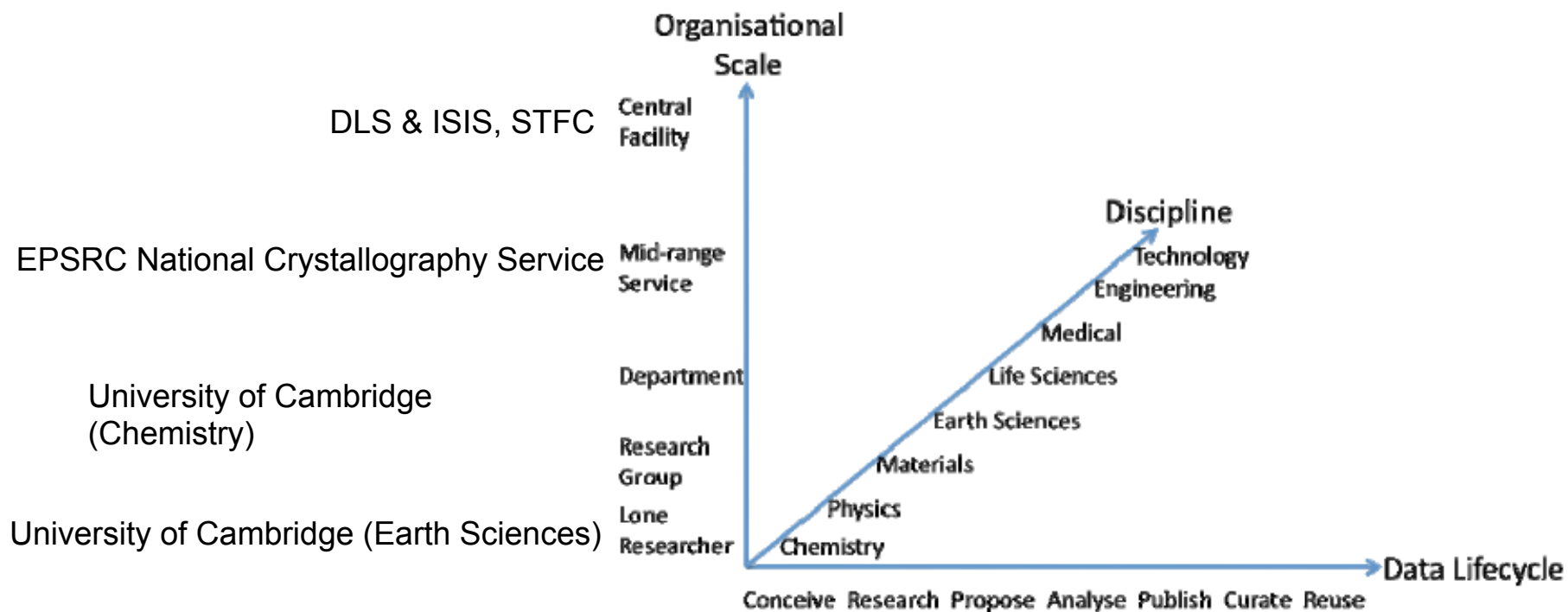
Infrastructure for
Integration in Structural Sciences

Objectives

Scale and complexity: small laboratory to institutional installation to large scale facilities e.g. DLS & ISIS, STFC

Interdisciplinary issues: research across domain boundaries

Data lifecycle: data flows and data transformations over time





Infrastructure for
Integration in Structural Sciences

Research Data & Infrastructure

- Research Data includes (all information relating to an experiment):
 - raw, reduced, derived and results data
 - research and experiment proposals
 - results of the peer-review process
 - laboratory notebooks
 - equipment configuration and calibration data
 - wikis and blogs
 - metadata (context, provenance etc.)
 - documentation for interpretation and understanding (semantics)
 - administrative and safety data
 - processing software and control parameters
- Infrastructure includes physical, technical, informational and human resources essential for researchers to undertake high-quality research:
 - Tools, Instrumentation, Computer systems and platforms, Software, Communication networks
 - Documentation and metadata
 - Technical support (both human and automated)
- Effective validation, reuse and repurposing of data requires
 - **Trust** and a thorough **understanding** of the data
 - **Transparent contextual and provenance information** detailing how the data were generated, processed, analysed and managed

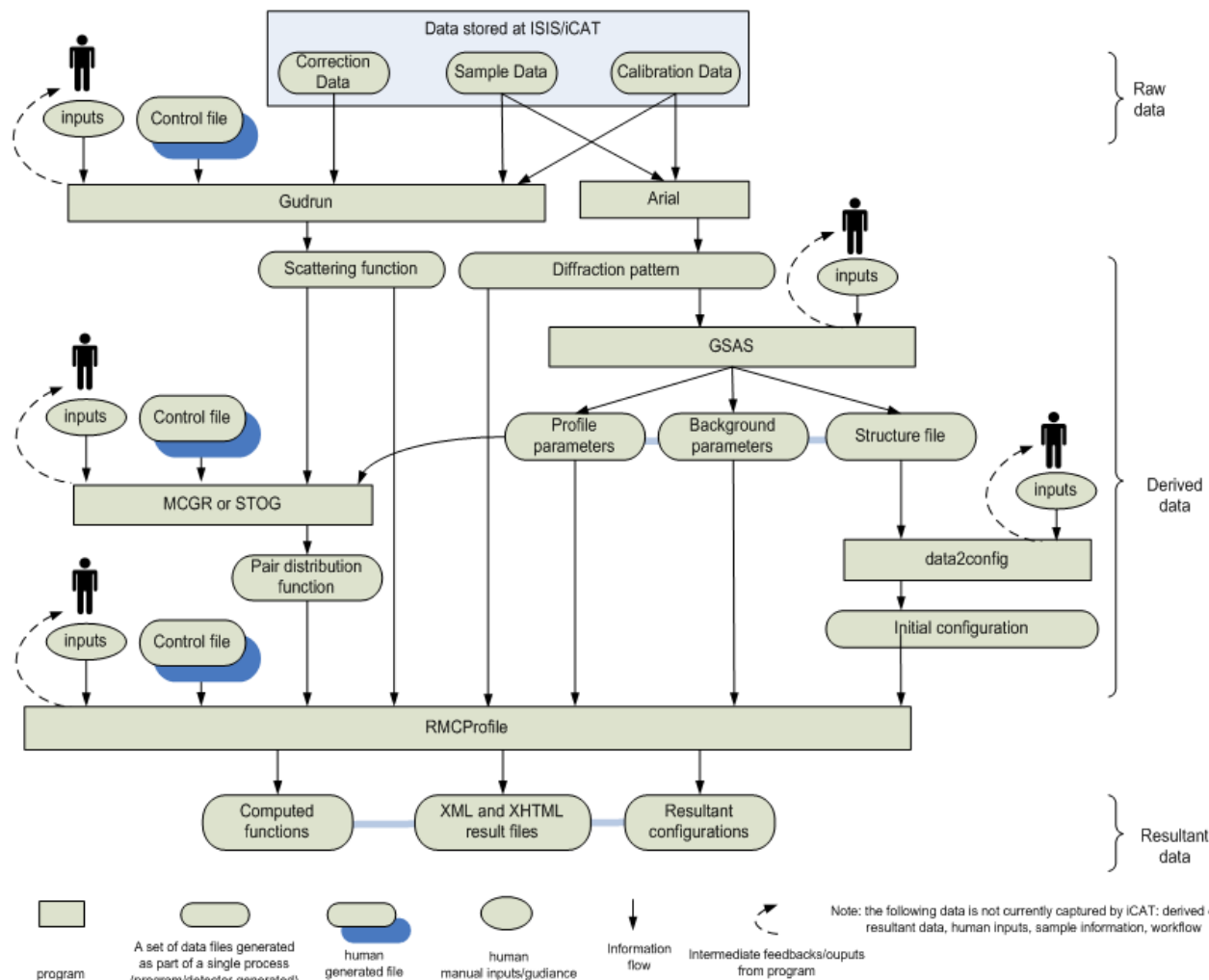


Infrastructure for
Integration in Structural Sciences

Earth Sciences, Cambridge

- Construct large scale atomic models of matter that best match experimental data; using Reverse Monte-Carlo Simulation techniques
- Experiment and data collection conducted at ISIS Neutron Source (GEM)
- Little or no shared infrastructure
 - Data sharing with colleagues via email, ftp, memory stick etc.
 - Data received from ISIS is currently stored on laptops or WebDAV server
- Management of intermediate, derived and results data a major issue
 - Data managed by individual researcher on own laptop
 - No departmental or central institutional facility
- Data management needs largely so that
 - Data can be shared internally
 - A researcher (or another team member) can return to and validate results in the future
 - External collaborators can access and use the data
- Any changes should be embedded into scientist's workflow and be non-intrusive

Earth Sciences, Cambridge: Typical workflow

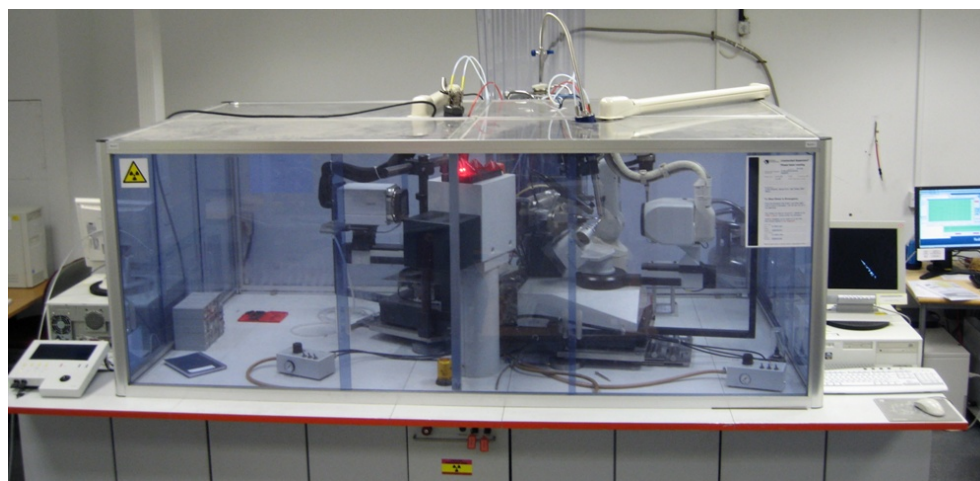




Infrastructure for
Integration in Structural Sciences

Chemistry, Cambridge

- Implementation and enhancement of a pilot repository for crystallography data underway (CLARION Project)
- Need for **IPR, embargo and access control** to facilitate the controlled release of scientific research data
- Information in **laboratory notebooks need to be shared** (ELN)
- Importance of **data formats and encodings** (RDF, CML) to maximise potential for data reuse and repurposing



EPSRC National Crystallography Service,
University of Southampton, UK



Infrastructure for
Integration in Structural Sciences

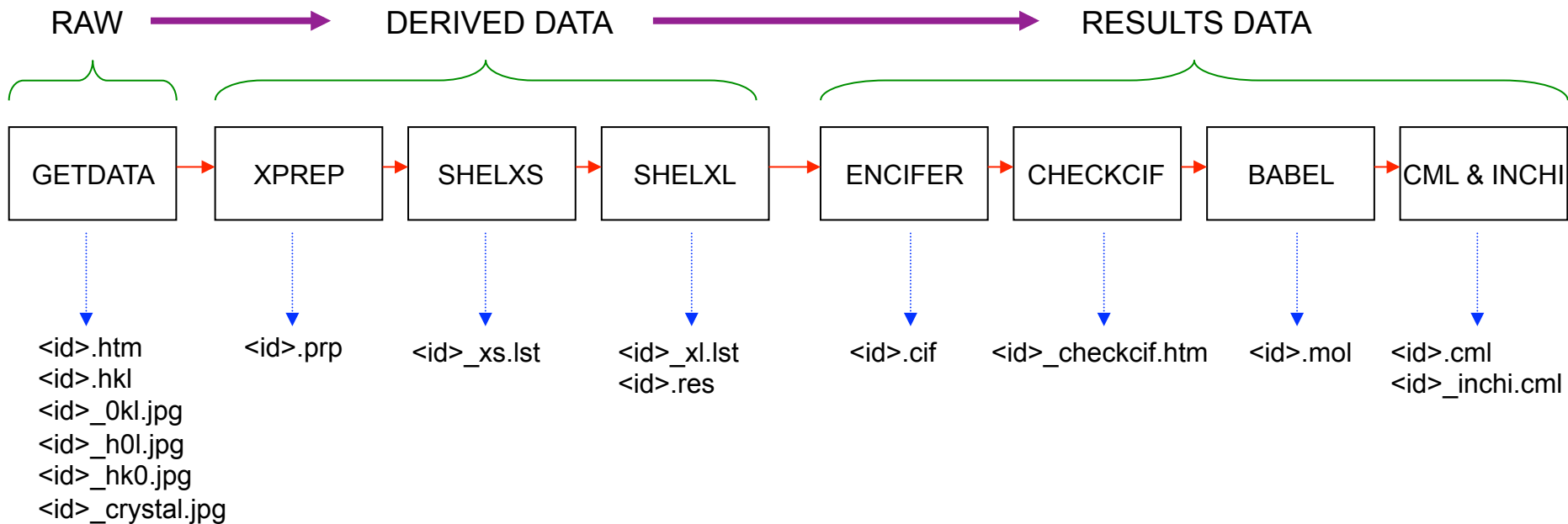
EPSRC NCS, Southampton

- **Service provision function** (operates nationally across institutions)
 - Local x-ray diffraction instruments + use of DLS (beamline I19)
 - Retain experiment data
 - Maintain administrative data
- Raw data generated in-house is stored at ATLAS Data Store (STFC)
- Local **institutional repository (eCrystals)** for intermediate, derived and results data
 - Metadata application profile
 - Public and private parts (embargo system)
 - Digital Object Identifier, InChi
- Experiments conducted and data collected by NCS scientists either in-house or at DLS
- **Labour-intensive paper-based administration and records-keeping**
 - Paper-based system for scheduling experiments
 - Paper copies of Experiment Risk Assessment (ERA) get annotated by scientist and photocopied several times
 - Several **identifiers** per sample
- **Administrative functions require streamlining between NCS and DLS**
 - e.g. standardisation of ERA forms, identifiers



Infrastructure for
Integration in Structural Sciences

EPSRC NCS: typical workflow



- **Initialisation**: mount new sample
- **Collection**: collect data
- **Processing**: process and correct images
- **Solution**: solve structures
- **Refinement**: refine structure

- **CIF**: produce Crystallographic Information File
- **Validation**: chemical & crystallographic checks
- **Report**: generate Crystal Structure Report
- **CML, INChI**



Infrastructure for
Integration in Structural Sciences

DLS & ISIS, STFC

- Operate on behalf of **multiple institutions and communities**
- Scientific (peer) and technical **review of proposals** for beam time allocation
- User offices manage **administrative and safety information**
- Service function implies an obligation to retain raw data
- Large infrastructure, engineered to **manage raw data**
 - Designed to describe facilities based experiments in Structural Science e.g. ISIS Neutron Source, Diamond Light Source.
 - ICAT implementation of Core Scientific Metadata Model (CSMD)
- **No storage or management of derived and results data**
 - Derived data taken off site on laptops, removable drives etc.
 - Results data independently worked up by individual researchers
- Experiment/Sample **identifiers based on beam line number**

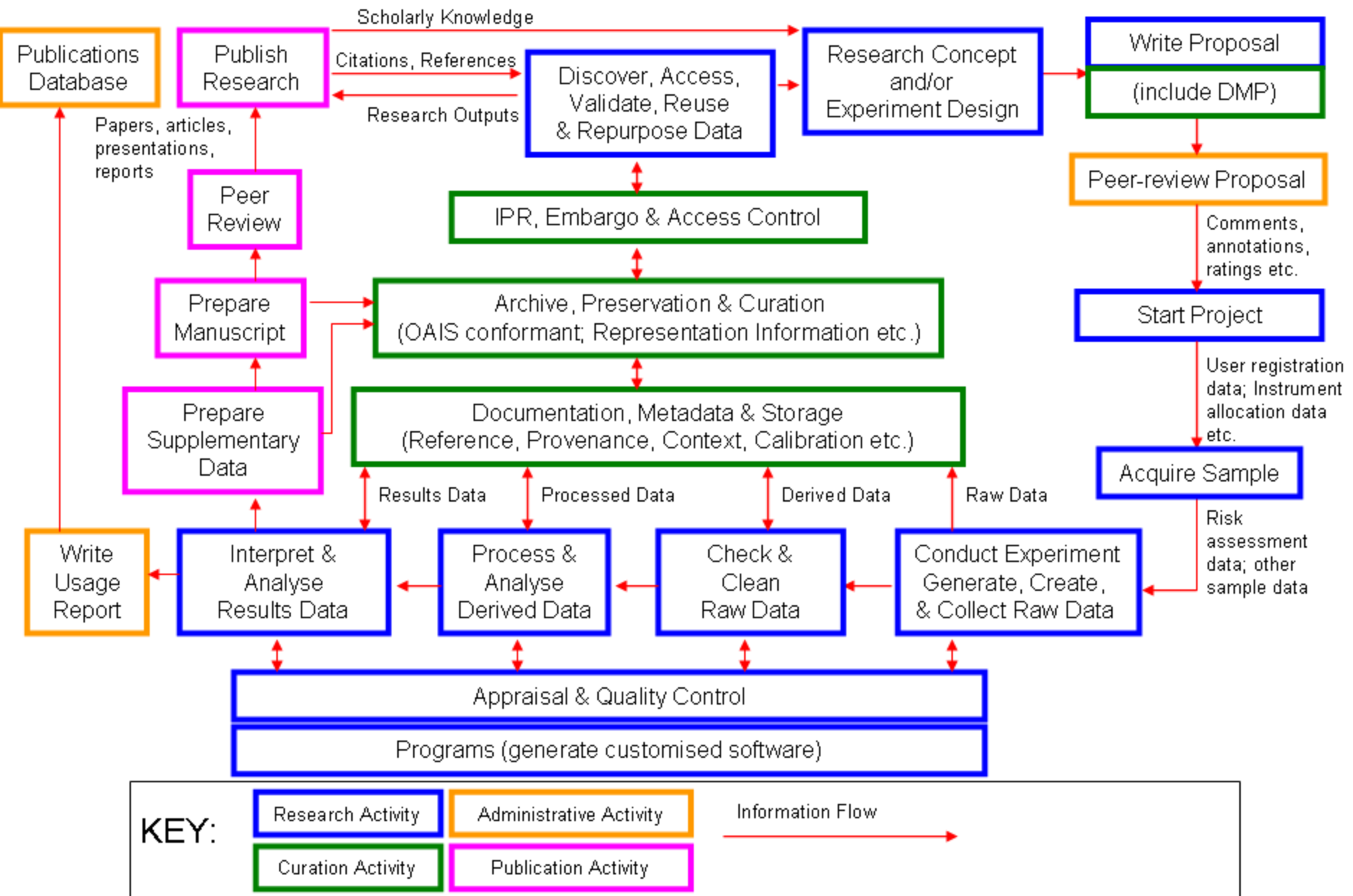


Infrastructure for
Integration in Structural Sciences

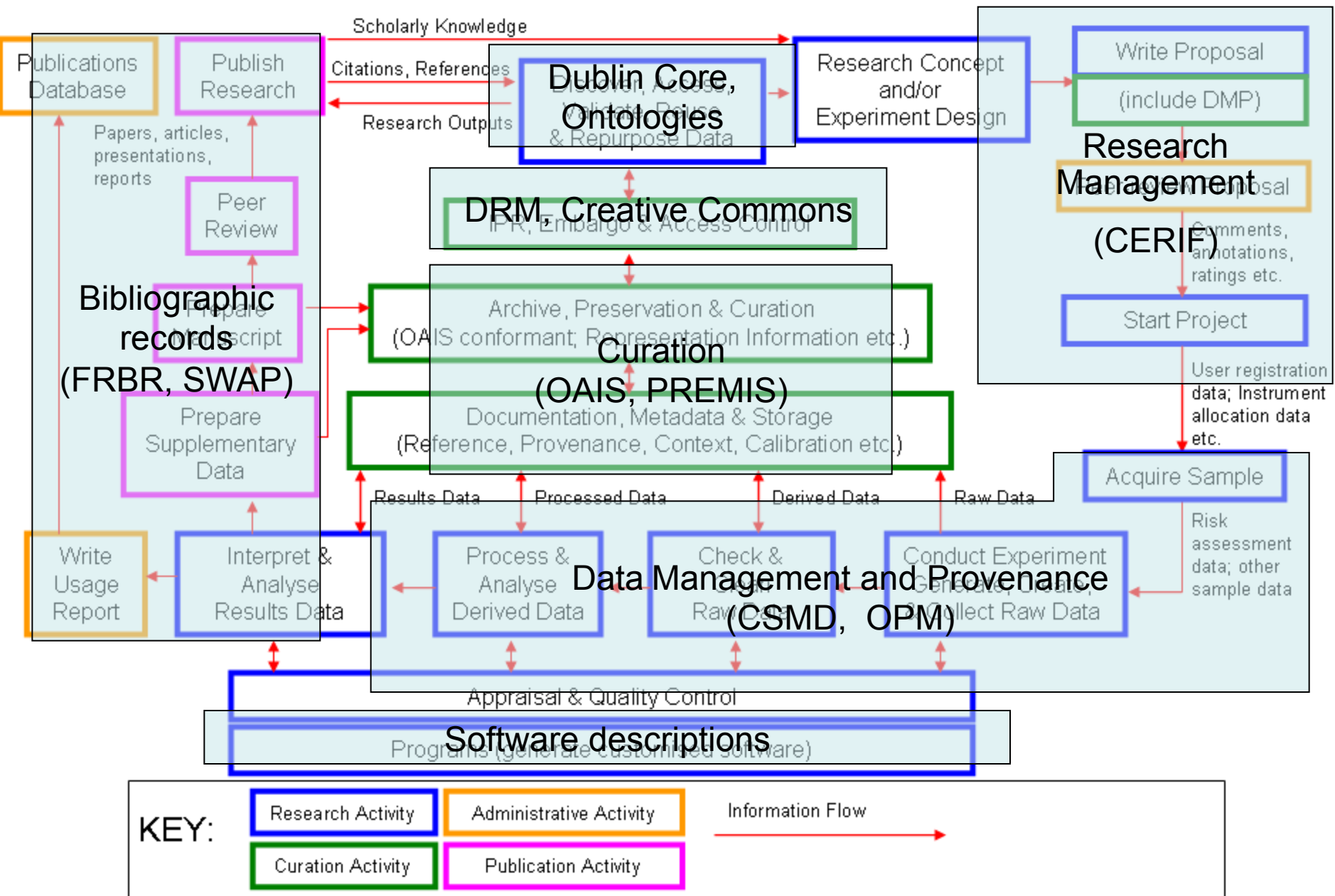
Generalised Requirements

- Basic requirement for **data storage and backup** facilities to sophisticated needs such as **structuring and linking** together of data
- **Contextual information is not routinely captured**
- The actual **workflow or processing pipeline is not recorded**
- Processing pipeline is dependent on a **suite of software**
- Adequate **metadata and contextual information** to support:
 - Maintenance and management
 - Linking together of all data associated with an experiment
 - Referencing and citation
 - Authenticity
 - Integrity
 - Provenance
 - Discovery, search and retrieval
 - Curation and preservation
 - IPR, embargo and access management
 - Interoperability and data exchange
- Simplification of inter-organisational communications and tracking, referencing and citation of datasets
 - **Standardised ERA forms**
 - **Unique persistent identifiers**
- Solutions should be as **non-intrusive** as possible

An Idealised Scientific Research Activity Lifecycle Model



An Idealised Scientific Research Activity Lifecycle Model

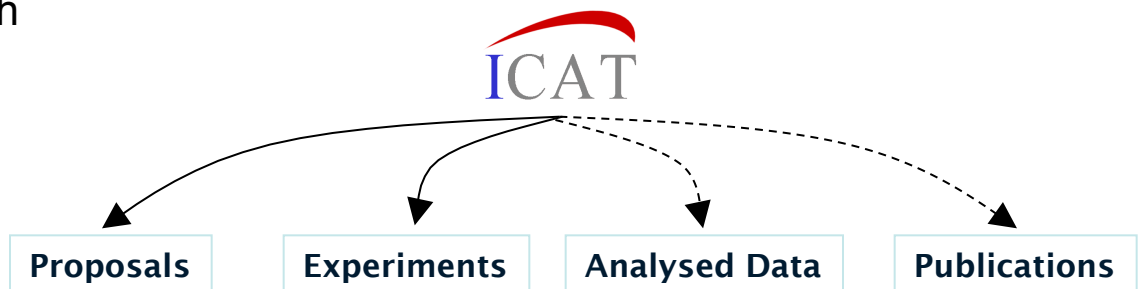
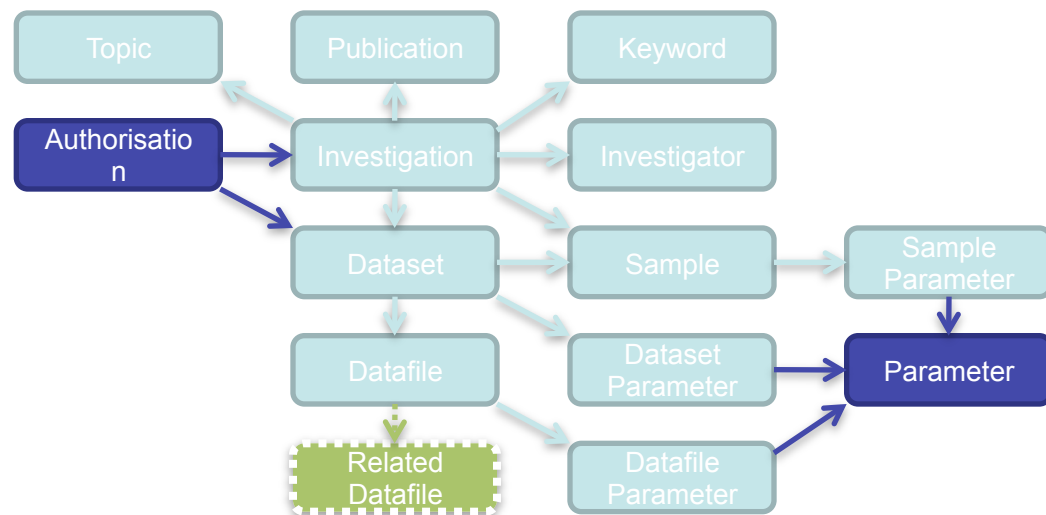




Infrastructure for
Integration in Structural Sciences

Core Scientific Metadata Model

- Designed to describe facilities based experiments in Structural Science
- CSMD is the basis of I2S2 integrated information model
- Forms a basis for extension to:
 - Laboratory based science
 - Derived data
 - Secondary analysis data
 - Preservation information
 - Publication data
- Aim to cover the scientist's research lifecycle as well as facilities data



<http://code.google.com/p/icatproject/>



Science & Technology
Facilities Council

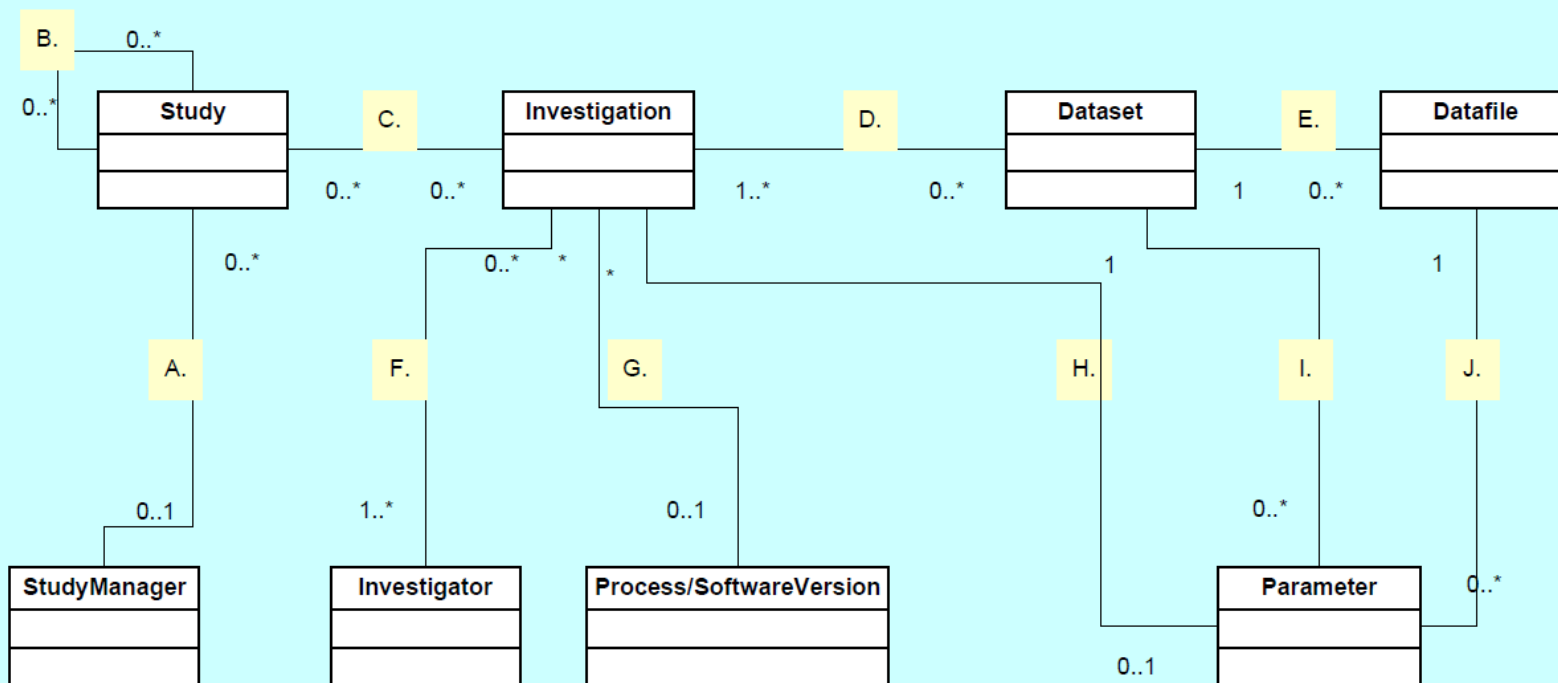


Infrastructure for
Integration in Structural Sciences

CSMD-Core

- Removal of facility specific information
- A simple model to describe datasets

I2S2 Core

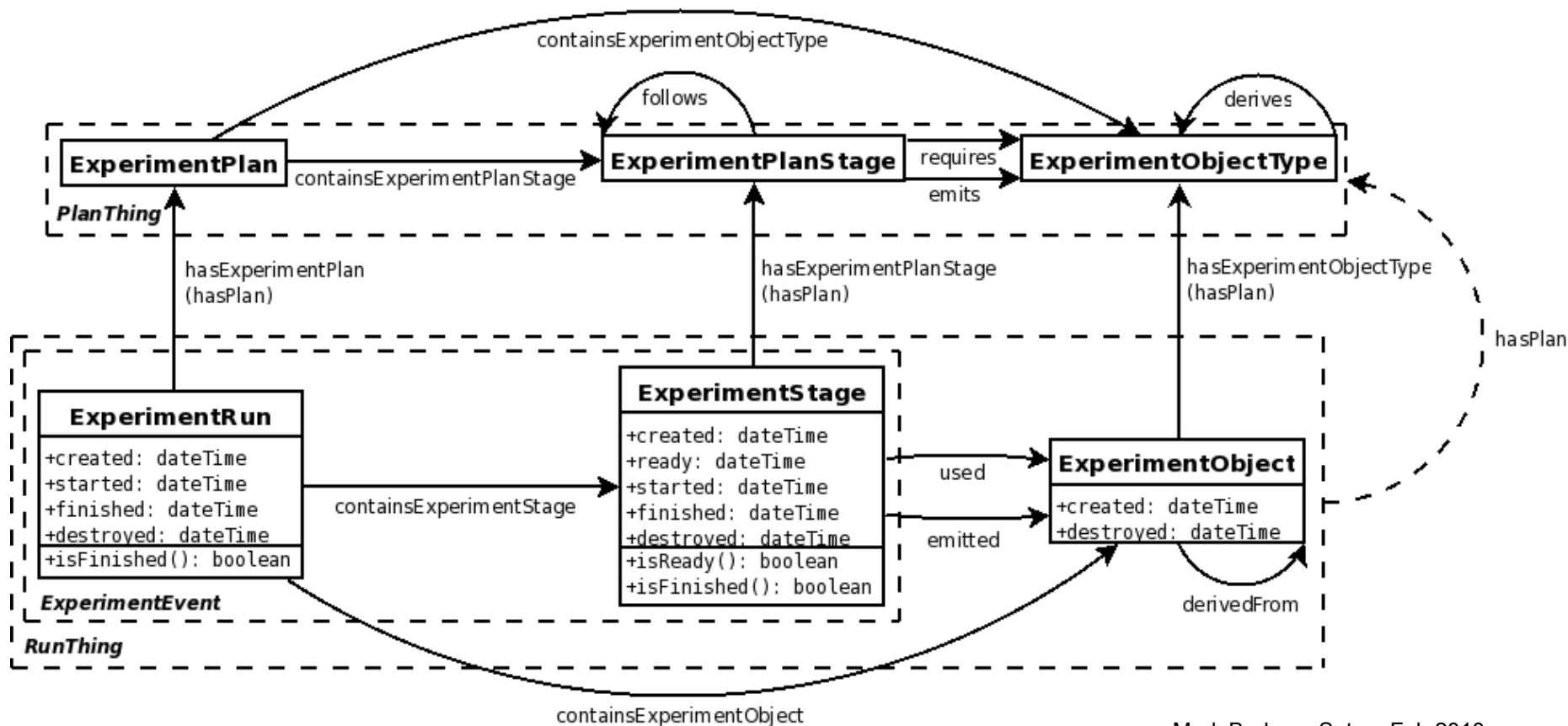




Infrastructure for
Integration in Structural Sciences

CoreChem Model

- An abstract model for **planning** and **enacting** chemistry experiments
- Enables exact **replication of methodology** in a machine-readable form
- Allows rigorous **verification of reported results**





Infrastructure for
Integration in Structural Sciences

I2S2 Integrated Information Model

- I2S2-IM = CSMD-Core + oreChem Model
- Use oreChem to describe **planning and enactment** of scientific process
- Use CSMD to **describe the data-sets** from the experiment
- I2S2-IM being implemented at STFC in the form of **ICAT-Lite**
 - A personal workbench for managing data flows
 - Allows the user to commit data
 - Enables capture of provenance information

A MS-viso version of Martin Dove's RMC workflow diagram
Last updated by Erica Yang, 05-01-2010

The diagram illustrates the RMC workflow, categorized into three main sections: Raw data, Derived data, and Resultant data.

Raw data: Includes "Data stored in ISIS/CAT" (containing "Reference Data", "Sample Data", and "Calibration Data") and "Inputs" (represented by a person icon).

Derived data: Includes "Guinix", "Aval", "Scattering function", "Diffraction pattern", "QIAS", "Profile parameters", "Background parameters", "Structure file", "dataconfig", and "Initial configuration".

Resultant data: Includes "RMCProfile", "Computed functions", "XML and XHTML result files", and "Resultant configurations".

Legend:

- program (represented by a box)
- A set of data files generated as part of a single process (represented by a rounded rectangle)
- human generated file (represented by a rounded rectangle with a blue border)
- human input (represented by a person icon)
- information flow (represented by a downward arrow)
- intermediate feedback loops from program (represented by a curved arrow)

Note: The following data is not currently captured by CAT: derived data, resultant data, human inputs, sample information, workflow.

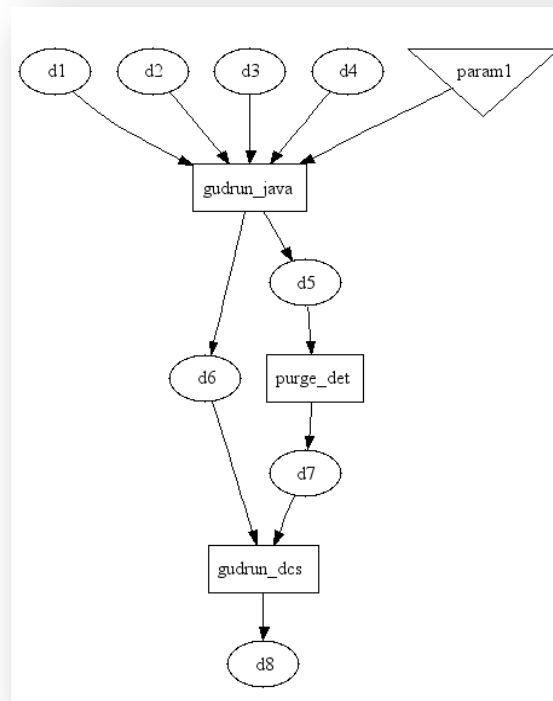
[illegible]

Yang_Eica (SIT\RCALISQ) > My Documents > EIC\SW > IS2 > Code > trunk > ICAT_LITE_CORE > test > GudmanFiles > run > SANDALS

Organize ▾ Include in library ▾ Share with ▾ Burn ▾ New folder ▾

Name	Date modified	Type	Size
dist	25-03-2010 11:19	NTFS File	1 KB
IS2 Demo	18-06-2010 10:49	Rating System File	1 KB
nbiobject	25-03-2010 11:19	RAWMION File	39 KB
src	25-03-2010 11:19	RAWTRAN File	39 KB
test	25-03-2010 11:19	SASRAT File	39 KB
svn	25-03-2010 11:19	SWMION File	39 KB
GudmanFiles_Eica	25-03-2010 11:19	TRANFIO File	2 KB
svn	25-03-2010 11:16	BAO File	12 KB
GudmanGUI2	18-06-2010 10:49	Microsoft Program...	5 KB
RunData	18-06-2010 10:49	Rating System File	1 KB
run	09-06-2010 11:42	BAO File	0 KB
svn	09-06-2010 11:42	Microsoft Program...	0 KB
SANDALS	09-06-2010 11:42	Rating System File	1 KB
svn	25-03-2010 11:19	RAWMION File	39 KB
Water	25-03-2010 11:19	NTFS File	1 KB
StartupFiles	25-03-2010 11:19	PLATE File	7 KB
icat	25-03-2010 11:19	RAWMION File	39 KB
testFolder	25-03-2010 11:19	RAWTRAN File	39 KB
xml-resources	25-03-2010 11:19	SWMION File	39 KB
ICAT_LITE_V5	25-03-2010 11:19	TRANFIO File	2 KB
svn	25-03-2010 11:19	VAIOAT File	24 KB
svn	18-06-2010 10:49	Microsoft Program...	5 KB
testfiles	18-06-2010 10:49	Rating System File	1 KB
Code\InfoAnalysis	25-03-2010 11:19	RAWMION File	39 KB
Data\ManagementTools	25-03-2010 11:19	RAWTRAN File	39 KB
Day2DayProgress	25-03-2010 11:19	SWMION File	39 KB
cd\resources	25-03-2010 11:19	BAO File	7 KB
graphics	25-03-2010 11:16	BAO File	51 KB
GridAP	25-03-2010 11:19	BAO File	45 KB
gridapp-winVn	25-03-2010 11:19	CHSIOU File	31 KB
IS2-activity-model	25-03-2010 11:19	DCSO File	568 KB
icatgoogle	25-03-2010 11:19	GR File	45 KB
information-model	25-03-2010 11:19	GR File	45 KB
meetings	25-03-2010 11:16	Microsoft Program...	4 KB
metadata	25-03-2010 11:19	GUD File	3 KB
Notes	25-03-2010 11:19	MGSIOU File	4 KB
proposal	25-03-2010 11:19	MBIOU File	4 KB
references	25-03-2010 11:19	MBIOU File	1 KB
requirements	25-03-2010 11:19	MBIOU File	1 KB
tools	25-03-2010 11:19	MBIOU File	1 KB

82 Items



Archive
Browse
Restore

Derived Data

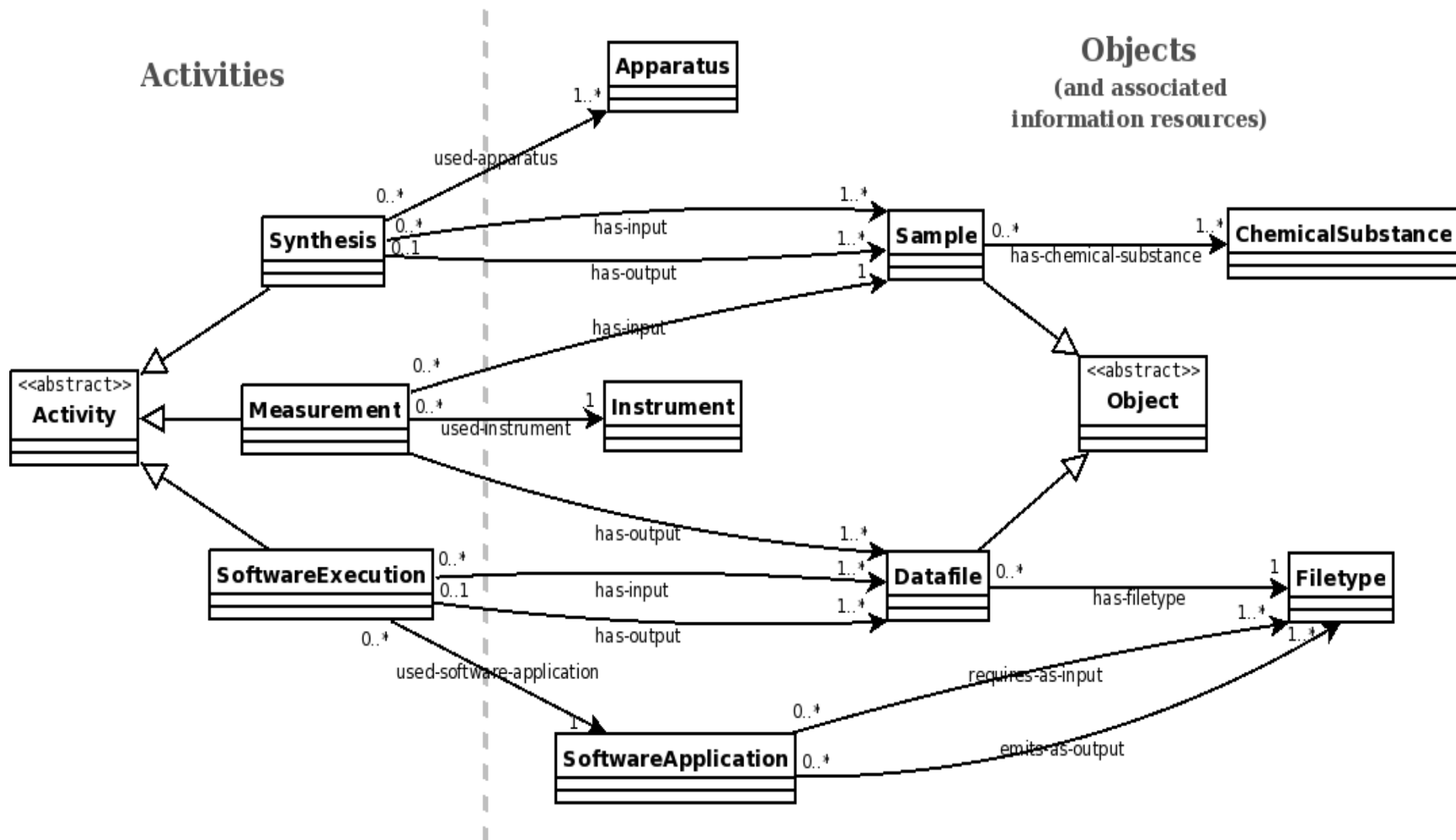


Infrastructure for
Integration in Structural Sciences

I2S2-IM

Activities

Objects (and associated information resources)





Infrastructure for
Integration in Structural Sciences

Testing the I2S2-IM

Case study 1: Scale and Complexity

- Data management issues **spanning organisational boundaries** in Chemistry
- Interactions between a lone worker or research group, the EPSRC NCS and DLS
- Traversing **administrative boundaries** between institutions and experiment service facilities
- Aim to probe both **cross-institutional and scale** issues

Case Study 2: Inter-disciplinary issues

- Collaborative group of inter-disciplinary scientists (university and central facility researchers) from both Chemistry and Earth Sciences
- Use of ISIS neutron facility and subsequent modelling of structures based on raw data
- Identification of **infrastructural components and workflow modelling**
- Aim to explore the **role of XML for data representation** to support easier sharing of information content of derived data



Infrastructure for
Integration in Structural Sciences

Cost-Benefits Analysis

- A **before and after cost-benefit analysis** using the Keeping Research Data Safe model
- **Extending the KRDS Model**
 - Focus has been on extensions and elaboration of activities in the research (KRDS “pre-Archive”) phase
- **Metrics and assigning costs**
 - Identification of activities in research activity lifecycle model that will represent significant cost savings or benefits
 - Work to identify non-cost benefits and possible metrics
- 2 use case studies
 - Quantitative -cost-benefits in terms of service efficiencies (NCS)
 - Qualitative -researcher benefits (improvement in tools; ease of making data accessible)



Infrastructure for
Integration in Structural Sciences

Conclusions

- Considerable variation in requirements between **differing scales of science**
- At present individual researcher, group, department, institution, facilities all **working within their own frameworks**
- Merit in adopting an **integrated approach** which caters for all scales of science:
 - Aggregation and/or cross-searching of related datasets
 - Efficient exchange, reuse and repurposing of data across disciplinary boundaries
 - Data mining to identify patterns or trends
- **I2S2 Integrated information model aims to:**
 - Support the scientific research activity lifecycle model
 - Capture processes and provenance information
 - Interoperate with and complement existing models and frameworks
- Before and after cost-benefits analysis to assess impact



Infrastructure for
Integration in Structural Sciences

Project Team

- Liz Lyon (UKOLN, University of Bath & Digital Curation Centre)
- Manjula Patel (UKOLN, University of Bath & Digital Curation Centre)
- Simon Coles (EPSRC National Crystallography Centre, University of Southampton)
- Brian Matthews (Science & Technology Facilities Council)
- Erica Yang (Science & Technology Facilities Council)
- Martin Dove (Earth Sciences, University of Cambridge)
- Peter Murray-Rust (Chemistry, University of Cambridge)
- Neil Beagrie (Charles Beagrie Ltd.)

m.patel@ukoln.ac.uk

<http://www.ukoln.ac.uk/projects/I2S2/>